

中文信息处理 Chinese Information Processing

第五章 作业

14307130318 刘超颖

1. 写程序处理布朗语料库，找到以下问题的答案：a. 哪些名词常以它们复数形式而不是它们的单数形式出现？（只考虑常规的复数形式，-s 后缀形式的）。b. 哪个词的不同词性标记数目最多？c. 按频率递减的顺序列出标记。前20 个最频繁的词性标记代表什么？d. 名词后面最常见的是哪些词性标记？这些标记代表什么？

a.

```
import nltk
from nltk.corpus import brown
wsj = nltk.corpus.brown.tagged_words(tagset = 'universal')
word_tag_fd = nltk.FreqDist(wsj)
sdist = [wt[0] for (wt, _) in word_tag_fd.most_common() if wt[1] =
= 'NOUN' and wt[0].endswith('s')]
scutdist = []
for w in sdist:
    w = w[:-1]
    scutdist.append(w)
result = []
word = brown.words()
fd = nltk.FreqDist(word)
for w in scutdist:
    if fd[w] < fd[w + 's']:
        result.append(w)
result
```

```
[u'year', u'State', u'eye', u'business', u'thing', u'member', u'word', u'Mis', u'  
student', u'proces', u'minute', u'mean', u'month', u'basi', u'condition', u'hour  
, u'mile', u'clas', u'term', u'Congres', u'friend', u'countrie', u'method', u's  
ale', u'serie', u"man'", u'arm', u'activitie', u'citie', u'leader', u'progres',  
u'element', u'analysi', u'factor', u'Jame', u'Thoma', u'event', u'facilitie', u'  
technique', u'dollar', u'stres', u'statu', u'Charle', u'glas', u'institution', u'  
'tree', u'product', u'studie', u'relation', u'ga', u'fund', u'clothe', u'succes'  
, u'inche', u'companie', u'los', u'Corp', u'requirement', u'circumstance', u'chu  
rche', u'parent', u'cell', u'mas', u'pres', u'worker', u'citizen', u'crisi', u'p  
oem', u'classe', u'feature', u'Loui', u'rule', u'resource', u'Jone', u'item', u'  
Texa', u'lip', u'adres', u'Pari', u'leg', u'finger', u'policie', u'employee', u'  
'politic', u'Lao', u'familie', u'bodie', u'aspect', u'yard', u'affair', u"one'",  
u'propertie', u'agencie', u'weapon', u'Negroe', u'storie', u'partie', u'emphasi  
, u'Jesu', u'Dalla', u'Nation', u'Lewi', u'processe', u'guest', u'song', u'head  
quarter', u'soldier', u'component', u'Jew', u'dres', u'flower', u'gras', u'troop  
, u'vehicle', u'Lo', u'wave', u'Massachusett', u'measurement', u'opportunitie',  
u'Angele', u'bond', u'centurie', u'stair', u'bird', u'difficultie', u'losse', u'  
'weaknes', u'panel', u'pound', u'shoe', u'sport', u'taxe', u'slave', u'qualitie'  
, u'Han', u'Clas', u'Motor', u'thicknes', u'emotion', u'darknes', u'possibilitie  
, u'expenditure', u'particle', u'acre', u'error', u'remark', u'communitie', u'n  
eighbor', u'Illinoi', u'Orlean', u'manufacturer', u'Democrat', u'atom', u'societ  
ie', u'customer', u'musician', u'knee', u'candidate', u"God'", u'arrangement', u'  
'Holme', u"father'", u'ear', u'specie', u'authoritie', u'Mari', u"year'", u'obse  
rvation', u'egg', u'bomb', u'expert', u'consequence', u'supplie', u"today'", u'o
```

b.

```
import nltk
from nltk.corpus import brown
words = nltk.corpus.brown.tagged_words()
cfd = nltk.ConditionalFreqDist(words)
ans = 0
for w in cfd.conditions():
    cnt = len(cfd[w].most_common())
    if cnt > ans:
        ans = cnt
        result = w
result
```

```
LCYmengmengdadeMacBook-Pro:中文信息处理 LCY$ python
Python 2.7.10 (default, Jul 30 2016, 18:31:42)
[GCC 4.2.1 Compatible Apple LLVM 8.0.0 (clang-800.0.34)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> from nltk.corpus import brown
>>> words = nltk.corpus.brown.tagged_words()
>>> cfd = nltk.ConditionalFreqDist(words)
>>> ans = 0
>>> for w in cfd.conditions():
...     cnt = len(cfd[w].most_common())
...     if cnt > ans:
...         ans = cnt
...         result = w
[...
>>> result
u'that'
>>> []
```

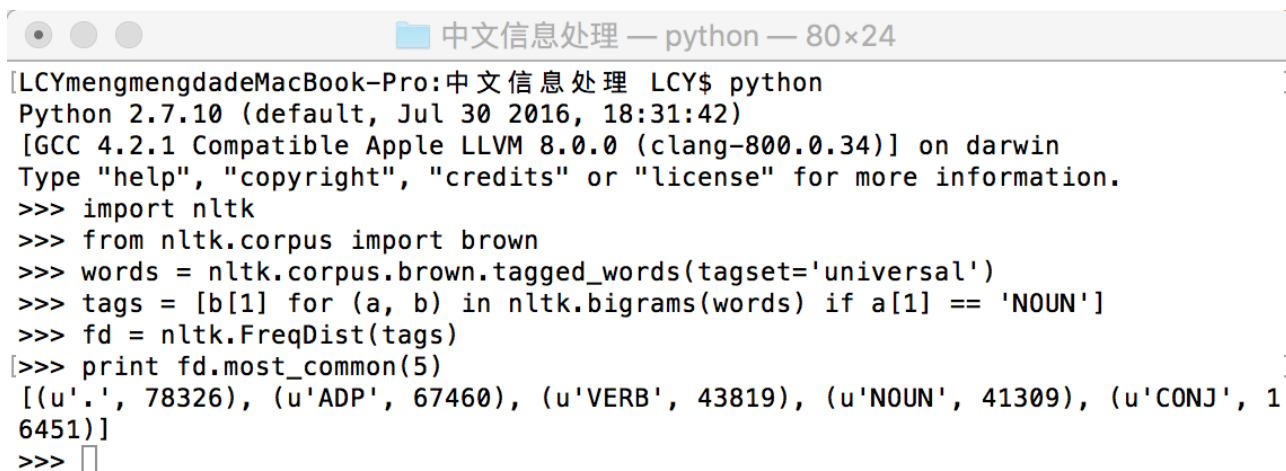
C.

```
import nltk
from nltk.corpus import brown
words = nltk.corpus.brown.tagged_words()
tag = []
for w in words:
    tag.append(w[1])
fd = nltk.FreqDist(tag)
print fd.most_common(20)
```

```
LCYmengmengdadeMacBook-Pro:中文信息处理 LCY$ python
Python 2.7.10 (default, Jul 30 2016, 18:31:42)
[GCC 4.2.1 Compatible Apple LLVM 8.0.0 (clang-800.0.34)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> from nltk.corpus import brown
>>> words = nltk.corpus.brown.tagged_words()
>>> tag = []
>>> for w in words:
...     tag.append(w[1])
[...
>>> fd = nltk.FreqDist(tag)
[>>> print fd.most_common(20)
[(u'NN', 152470), (u'IN', 120557), (u'AT', 97959), (u'JJ', 64028), (u'.', 60638)
, (u',', 58156), (u'NNS', 55110), (u'CC', 37718), (u'RB', 36464), (u'NP', 34476)
, (u'VB', 33693), (u'VBN', 29186), (u'VBD', 26167), (u'CS', 22143), (u'PPS', 182
53), (u'VBG', 17893), (u'PP$', 16872), (u'TO', 14918), (u'PPSS', 13802), (u'CD',
13510)]
>>> ]
```

d.

```
import nltk
from nltk.corpus import brown
words = nltk.corpus.brown.tagged_words(tagset='universal')
tags = [b[1] for (a, b) in nltk.bigrams(words) if a[1] == 'NOUN']
fd = nltk.FreqDist(tags)
print fd.most_common(5)
```



```
LCYmengmengdadeMacBook-Pro:中文信息处理 LCY$ python
Python 2.7.10 (default, Jul 30 2016, 18:31:42)
[GCC 4.2.1 Compatible Apple LLVM 8.0.0 (clang-800.0.34)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> from nltk.corpus import brown
>>> words = nltk.corpus.brown.tagged_words(tagset='universal')
>>> tags = [b[1] for (a, b) in nltk.bigrams(words) if a[1] == 'NOUN']
>>> fd = nltk.FreqDist(tags)
>>> print fd.most_common(5)
[(u'.', 78326), (u'ADP', 67460), (u'VERB', 43819), (u'NOUN', 41309), (u'CONJ', 16451)]
>>>
```

2. 讲义中绘制曲线显示了查找标注器的性能随模型的大小增加的变化。请仿照该方法绘制当训练数据量变化时unigram标注器的性能曲线。

```
import nltk
from nltk.corpus import brown

def performance(size):
    brown_tagged_sents = brown.tagged_sents(categories = 'news')
    train_tagged_sents = brown_tagged_sents[:size]
    unigram_tagger = nltk.UnigramTagger(train_tagged_sents)
    return unigram_tagger.evaluate(brown_tagged_sents)

def display():
    import pylab
    sizes = [1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000,
10000]
    perfs = []
    for s in sizes:
        perfs.append(performance(s))
    pylab.plot(sizes, perfs, '-bo')
    pylab.title('Unigram Tagger Performance with Varying Model Size')
    pylab.xlabel('Model Size')
    pylab.ylabel('Performance')
    pylab.show()

>>> import evaldisplay
>>> evaldisplay.display()
```

Figure 1

